STATE OF ILLINOIS

ILLINOIS COMMERCE COMMISSION

Illinois Bell Telephone Company, AT&T Communications of Illinois, Inc., TCG Illinois, TCG Chicago, TCG St. Louis, WorldCom, Inc., McLeodUSA Telecommunications Services, Inc., XO Illinois, Inc., NorthPoint Communications, Inc., Rhythms Netconnection and Rhythms Links, Inc., Sprint Communications L.P., Focal Communications Corporation of Illinois, And Gabriel Communications of Illinois Inc. Petition for Resolution of Disputed Issues Pursuant to Condition (30) of the SBC/Ameritech Merger Order.	)	No. 01-0120

DIRECT TESTIMONY

OF

DR. DANIEL S. LEVY

On Behalf of

AMERITECH ILLINOIS

July 13, 2001

DIRECT TESTIMONY
OF DR. DANIEL S. LEVY
ON BEHALF OF AMERITECH ILLINOIS

I. BACKGROUND, QUALIFICATIONS, AND PURPOSE OF TESTIMONY.

Q. PLEASE STATE YOUR NAME AND BUSINESS ADDRESS.

A. My name is Dr. Daniel S. Levy. My business address is 33 West Monroe Street, Chicago, Illinois.

Q. BY WHOM ARE YOU EMPLOYED AND IN WHAT CAPACITY?

A. I am a Partner at Arthur Andersen, LLP, where I serve as the National Director of Economic Consulting for Arthur Andersen's Business Consulting Group. In that capacity, I advise clients as to the use of statistical analysis and techniques in business and in judicial and regulatory proceedings.

Q. WHAT IS THE PURPOSE OF YOUR TESTIMONY?

A. In this proceeding, the Commission will decide whether Ameritech Illinois should keep its current plan for paying remedies to competing local exchange carriers ("CLECs") in the event it fails to meet certain performance standards, or whether to replace the current remedy plan with one proposed by the CLECs. Each plan has a different methodology for using statistical analysis to test compliance with performance standards. The purpose of my testimony is to explain and compare the statistical methodologies used in the Ameritech Illinois and CLEC plans, and to show why Ameritech Illinois' methodology is balanced, practical, scientifically valid, and consistent with the goals of the Telecommunications Act of 1996 ("1996 Act").

Q. PLEASE DESCRIBE YOUR BACKGROUND AND QUALIFICATIONS FOR REACHING THAT CONCLUSION.

A. I have a PhD. in Economics from The University of Chicago. I have over 20 years of experience in research and advising clients, particularly on matters related to statistics. And for more than three years, I have advised Ameritech Illinois and its affiliates with respect to the implementation and ongoing administration of performance remedy plans in general, and the remedy plan that is used in Illinois in particular.

Q. ARE YOU ALSO FAMILIAR WITH THE CLEC REMEDY PLAN?

A. Yes. I have reviewed the plan that the CLECs filed in this docket on March 12, 2001. I have also reviewed virtually identical plans that these CLECs submitted in Michigan, Indiana, Ohio, and Wisconsin, and I have reviewed testimony by CLEC witnesses regarding those plans. Further, I have reviewed similar statistical methodologies proposed by CLECs for use in the ongoing third-party tests of operations support systems ("OSS") in the Ameritech states.

Q. IN ADDITION TO YOUR WORK AT ARTHUR ANDERSEN, WHAT POSITIONS HAVE YOU HELD?

A. Prior to joining Arthur Andersen, I performed research and consulting work for Needham-Harper Worldwide Advertising, The University of Chicago Computation Center, SPSS Inc., The RAND Corporation, and Charles River Associates. I joined Arthur Andersen in 1995 as an economist. I became the Regional Director of Economics for Arthur Andersen's Central Region (which includes Illinois) one year later. I was promoted to National Director in 1998, and have served in that position since. My resume is attached hereto as Attachment A.

Q. WHAT IS THE PURPOSE OF STATISTICAL ANALYSIS?

A. Generally speaking, the goal of statistics is to analyze and interpret data and to objectively determine the reliability of the conclusions. These methods can be, and are, applied to almost every facet of everyday life. For example, one can look at the results of surveys to test and draw conclusions about public opinion, or at economic data to draw conclusions about the effects of an event or decision, or at the results of a scientific experiment to test whether a theory is correct. Or one can flip a coin to test whether it is fair. Statistical methods are often used to determine whether some factor, or factors, cause two populations to differ. For example a school system may use statistical methods to determine whether a given teacher produces a population of students that perform better on standardized tests than a population of students taught by other teachers. Each of these examples uses observed data to test a hypothesis and to draw conclusions about a population or populations.

To draw valid conclusions, one must address the reality that individual observations of data are often subject to random variation. On average, a fair coin will come up "heads" 50 percent of the time and "tails" the other 50 percent. If you could flip the coin an infinite number of times, you would tend to see results that are closer to an equal number of heads and tails. But in real world settings we do not have an infinite amount of data. If you flip a coin only twice, there is a 50 percent chance that it will come up one head and one tail, but there is also a 50 percent chance that it will produce either two heads or two tails. The coin may still be fair if it produces two heads or two tails; it's just that you only looked at two flips of the coin instead of a very large number. Thus, before jumping to a conclusion that the coin is unfair because it came up heads two

straight times, you need to consider the impact of random variation. That is also why survey results generally come with a margin of error. Statistical analysis provides a scientific method for factoring that random variation into the thought process.

Q. HOW DO STATISTICAL PRINCIPLES APPLY TO PERFORMANCE MEASUREMENT, STANDARDS, AND REMEDIES?

A. Performance measurement is just one of the many applications in which observed data is used to test a hypothesis about the population. In the case of performance measurement, Ameritech Illinois reports data about its performance of numerous functions. The data are reported separately for each CLEC, for a number of different products or services, for designated geographic areas. Ameritech Illinois compares each of these performance results to a standard. These standards are based on the principle of nondiscrimination or "parity" that is set forth in the 1996 Telecommunications Act, and in the criteria established by this Commission and the FCC. In most cases, the standard is "parity" between the wholesale function and a retail analog. In some cases, there is no retail analog, and the standard is set by a "benchmark": for example, 99 percent of mechanized completions are to be returned within 1 hour.

You can think of each of these performance measurements as a "test," the purpose of which is to draw a conclusion about whether Ameritech Illinois is satisfying its obligation to provide nondiscriminatory service, and whether Ameritech Illinois' personnel, electronic systems, and procedures are functioning in a nondiscriminatory manner. As Mr. Fioretti describes in his affidavit, there are over 160 performance measures, which are further divided into thousands of product, service, and geographic categories.

As with data used in other tests, performance data are subject to random variation. For example, in December 2000 (simulated data), the average time required for Ameritech Illinois to install retail residential POTS not requiring a field visit in the Chicago geographic region service was 0.65 days. But obviously each and every installation would not take exactly 0.65 days. Instead, some installations take less time, while others take more. Thus, if you randomly picked some installations out of the monthly total, the average time for those installations would likely be different from the overall average. Similarly, if you looked at the average installation time for CLEC customers, it too would likely be different from the overall retail average, even though CLEC customers are receiving the same level of service as Ameritech's own retail customers.

Q. WHAT CAUSES RANDOM VARIATION IN PERFORMANCE DATA?

A. There are an almost infinite number of causes. For example, one installation might take longer than another because of weather, or traffic conditions, or because the installation itself is more or less complicated or difficult to complete. These random events will cause the observed level of service provided to the CLEC's customers to appear better than that provided to Ameritech Illinois customers in some months and worse in other months, simply due to indiscriminant random variation.

Q. WHY IS IT IMPORTANT TO ADDRESS SUCH RANDOM VARIATION IN A PERFORMANCE REMEDY PLAN?

A. A remedy plan is designed to enforce performance standards and the underlying requirement of nondiscrimination. The basic idea is that, if Ameritech Illinois is discriminating against a CLEC or CLECs, it will pay a remedy to that CLEC (or to the

State). This provides compensation to the affected parties, and it provides an incentive to Ameritech Illinois to behave in a nondiscriminatory fashion.

If a remedy plan is to really serve its purpose, it should require remedies only where discrimination has really occurred. As I said earlier, individual performance observations (and thus average performance for different groups or samples of observations) are subject to random variation. Thus, even if there is no discrimination, you will see a difference in the average performance between two groups of performance metrics data. In fact, as I noted above, you will see a difference in performance between two randomly selected groups taken entirely from Ameritech Illinois retail data, even though Ameritech Illinois by definition does not discriminate against itself.

A remedy plan that forces Ameritech Illinois to pay remedies when it does *not* discriminate will not create the proper incentives. It would be like making parking meters with random timers on them and then giving motorists parking tickets when the meters randomly expired. It is simply an arbitrary and capricious transfer of funds from one company to another that provides no benefit to consumers or enhancement to competition. The payment of remedies when the underlying level of service is in parity reduces Ameritech Illinois' incentive to provide parity of service. Furthermore, as discussed in more detail below, remedy payments when parity exists will inhibit effective competition in the market and reduce the incentive for Ameritech Illinois to introduce new products and technology. The purpose of statistical analysis is to account for random variation and thus increase the chance that when remedies are paid they result from actual disparity in service.

Q. HOW DO THE COMPETING REMEDY PLANS ADDRESS THIS ISSUE?

A. The Ameritech Illinois and CLEC remedy plans both recognize the problem of random variation, but they take very different approaches to address it. In Section II, I describe the statistical methodology that Ameritech Illinois uses in its current remedy plan. In Section III, I contrast the partially developed methodology that the CLECs have developed to date, and show why Ameritech Illinois' plan is preferable. Section IV illustrates how the respective remedy plans work, by using simulated performance data for a three-month period.

II. DESCRIPTION OF STATISTICAL METHODOLOGY IN AMERITECH ILLINOIS' REMEDY PLAN.

Q. PLEASE OUTLINE THE APPROACH THAT AMERITECH ILLINOIS TAKES IN ITS REMEDY PLAN TO ADDRESS RANDOM VARIATION.

A. Ameritech Illinois' remedy plan follows the general approach the CLECs originally developed. The test that the CLECs proposed was a "Z-test."[1]

The basic idea of the Z-test is to consider the size of the difference between observed performance and the applicable standard to determine whether the difference is larger than what would often be found due to simple random variation in the data. The larger the difference, the more likely it is that there is some underlying disparity in performance as opposed to some observed difference that happened by random chance. For example, if retail repairs take 24 hours, one is more likely to find a real disparity in

[1] The test for interval measures is modified from the standard Z-test found in statistical texts. Throughout my testimony, I will generically refer to both the modified Z-test and the standard Z-test as the "Z-test."

performance if wholesale repairs take 240 hours than if wholesale repairs take 24.001 hours. Or, one is more likely to decide that a coin is unfair if it turns up "heads" 50 out of 50 times, than if it comes up heads 26 out of 50 times. Statistical science provides a rigorous way of looking at the difference, and the extent of the random variation in the data to figure out how likely it is that the difference is due to some real disparity as opposed to random chance – in other words, that the difference is "statistically significant." Ameritech Illinois' plan uses tests designed to achieve a 95 percent confidence level; roughly speaking, this means that an apparent shortfall in performance is considered "statistically significant" if the odds are 95 percent that it is caused by behavior rather than random chance.

Of course, that means the odds are still 5 percent that the apparent shortfall is caused by random chance, and with a large number of performance tests, a 5 percent error rate virtually guarantees that some measures will appear disparate even when they are in parity. With the introduction of remedy payments, this means that Ameritech Illinois would pay remedies on 5 percent of all tests performed even when the level of service to CLEC customers and Ameritech Illinois retail customers are in perfect parity. Therefore, as an integral part of achieving the CLECs' 95 percent confidence level, the CLECs proposed (and Ameritech Illinois has adopted) a standard statistical technique that considers the results of all the individual tests in the aggregate. The principle here is that, at the aggregate level, one is more likely to find a real disparity in performance if 100 out of 100 individual tests suggest a "failure" or disparity, than if only 1 out of 100 tests suggests disparity. In fact, given a 5 percent error rate, you would expect a few individual tests to suggest disparity based solely on random variation rather than any real

disparity in behavior. Again, statistical analysis provides a scientific method for determining how many individual test "failures" are expected due to random chance, and how many failures would be necessary to suggest a real disparity. The number of such failures is called "k" and the method for calculating that number is the "K table."

This overall integrated method of testing for differences between groups across a large number of tests is based on statistical techniques that are commonly accepted in the field of statistical research. They were developed for use in performance testing by the CLECs, and they have been approved by state commissions and the FCC. It is these very tests that form the basis of the Ameritech Illinois remedy plan.

Q. PLEASE TELL US IN MORE DETAIL ABOUT THE INDIVIDUAL PERFORMANCE TESTS AMERITECH ILLINOIS USES.

A. As I said earlier, most of Ameritech Illinois' wholesale performance standards are based on a comparison to retail performance. For example, one measure compares the average time to repair the CLEC's resale residential POTS service to the average repair interval for Ameritech Illinois' retail residential POTS service. I call these tests "parity" tests. In performance testing, we are checking to see if there is disparity between retail and wholesale performance for the month. (I describe the other kinds of tests, "benchmark" tests, later on in my testimony.) More specifically, we are trying to determine with a reasonable degree of scientific certainty whether retail performance is *better* than wholesale performance, because we are trying to guard against the possibility that Ameritech Illinois would treat itself better than it treats CLECs. The remedy plan is not concerned with situations in which retail performance is worse than wholesale, and Ameritech Illinois does not receive a credit in those situations. Because we are

concerned with potential disparities only in one direction (worse than retail performance), we describe the test as a one-tailed test.

Q. HOW DOES AMERITECH ILLINOIS TEST PARITY BETWEEN RETAIL AND WHOLESALE PERFORMANCE?

A. As I described above, the objective is to look at the difference between retail and wholesale performance, and figure out whether the difference falls within the range we would expect due to mere random variation, or whether it is large enough to suggest some underlying disparity in performance. For example, in December 2000 (simulated data) it took Ameritech Illinois 0.45 days, on average, to install wholesale residential POTS without fieldwork in Chicago. There were 8,792 such installations; some took more time than average, some less. The average interval for one CLEC (code-named "146" to protect its identity) that month was 0.55 days; there were 496 installations, some longer, some shorter than average, the same as with retail. Because the individual observations are different, the difference between the retail and wholesale averages may simply be the product of random chance: which carrier happened to get more of the shorter installations that month.

Q. HOW DO YOU DECIDE WHETHER THE DIFFERENCE IS DUE TO RANDOM CHANCE?

A. Just as there are commonly accepted ways to measure distance (feet, miles, meters) or time (minutes, hours, etc.), statistical science provides a number of standard ways to measure the degree of variation in data and to estimate the range of random variation we would expect between two samples of such data. Here, we look at the degree of variation in the retail data to determine the amount of variation that would be expected due to

random chance. Differences between retail observations are, almost by definition, due to random variation rather than discrimination because Ameritech Illinois would not discriminate against itself. Where there are at least 30 CLEC observations for interval measures, we measure the degree of variation by using a statistical measure called the "modified Z statistic." (As I describe below, there is an alternative statistical measure that works better when there are less than 30 observations.)

1. Parity Tests: Sample Size Of 30 Or More

Q. WHAT IS THE SOURCE OF THE MODIFIED Z STATISTIC?

A. It was developed and proposed by a group of carriers – AT&T, MCI (now WorldCom), Sprint, and LCI – that called themselves the Local Competition User Group or "LCUG." They first advanced the modified Z-statistic in 1998, during the FCC's rulemaking on performance measurement. Over time, a consensus developed as other carriers agreed to LCUG's approach. Ameritech Illinois accepted the modified Z test in the spirit of compromise, in order to allow statistical testing to commence.

Q. WHY IS IT CALLED A "MODIFIED" Z STATISTIC?

A. The modified Z-statistic is based on a commonly used statistical measure known as the "Z-statistic," which is designed to assess whether an observed difference between two averages is statistically significant. The LCUG worried that the standard Z-test, while testing for differences in the average service provided to CLEC and Ameritech Illinois retail customers, did not test for differences in the consistency or "variance" of service provided to CLEC customers. The LCUG suggested that incumbent carriers would provide their own retail customers and CLEC customers with the same average service, but could achieve this same average level of service to the CLEC customers by providing

some CLEC customers with very high quality service and other CLEC customers with very low quality or slow service. The LCUG suggested that this type of increased variation in level of service to CLEC customers would not only constitute a lower quality of service in itself, but that it would also reduce the effectiveness of a Z-test to detect any potential differences in the average level of service provided to CLEC and Ameritech Illinois customers. The LCUG, therefore, proposed the "modified" Z-test, which substitutes the incumbent LEC ("ILEC") standard deviation for the CLEC standard deviation in the standard Z formula. The benefit of this test is that differences in the variation in service provided to CLEC customers would not reduce the effectiveness of the test in detecting differences in the average level of service provided to CLEC and Ameritech Illinois customers.

For metrics that are measured as rates and proportions, the average performance determines the variance. It is not possible to maintain the same average performance while increasing the variance in performance of a rate or a proportion. Therefore for rates and proportions there is no need to modify the standard Z-test.

Q. WHY IS THE MODIFIED Z TEST NOT USED FOR PROPORTIONS AND RATES?

A. It is not necessary. As I stated above, the reason for using the modified Z-test for intervals is because there is concern that the ILEC would have an incentive to inflate the variability of the CLEC performance in order to pass the parity test while still providing better service to itself. For rates and proportions, the ILEC is unable to pursue a strategy of increasing the variance of the data in order to achieve a lower Z-score, because for rates and proportions the variability (variance) cannot be increased without also altering

the proportion or rate itself. Therefore, the more conventional pooled Z-test is used for these measures.

Q. HOW ARE THE Z-STATISTICS CALCULATED?

A. The Z-statistic is a commonly accepted statistical tool that uses the mean of the data and a commonly accepted measure of variation called the "standard deviation," which measures the normal or "standard" amount by which the individual data observations differ or "deviate" from the overall average. The Greek letter sigma (σ) is a shorthand symbol for the standard deviation. There are slightly different formulas for computing the Z-statistic, depending on whether the performance measurement is an average (the average time to repair service), a percentage (the percentage of due dates missed), or a rate (the rate of trouble reports). These formulas are well known in the field, and they are illustrated in section 3.0 of the Ameritech Illinois remedy plan. Attachment B to my testimony illustrates the calculation of a modified Z-statistic with a numerical example.

Q. AFTER CALCULATING "Z," WHAT IS THE NEXT STEP IN TESTING PARITY?

A. The next step is to see if Z, the measure of difference between average wholesale and retail performance, falls within the range of differences we would expect due to random chance. We do that by comparing the modified Z-statistic to the amount of difference one would expect from random variation, which is called the "critical" Z value.

Q. HOW DOES AMERITECH ILLINOIS CALCULATE THE CRITICAL Z VALUE?

A. Using standard statistical methods, AT&T developed a table that lists the combination of critical Z-values, and the number of apparently disparate test results, that would indicate true underlying disparity at a 95 percent confidence level. In other words, if the test indicates a failure or disparity, there is 95 percent confidence that there was a real disparity. Conversely, you can say that there is a 5 percent risk that the test will indicate disparity (in error) where there is none. These false alarms are referred to as "Type I" errors.

To determine the relevant critical Z-value under the Ameritech Illinois plan, all you need to know is the number of parity tests that are performed for a given CLEC. Once this is known you simply look up the combination of the critical Z-value and the number of missed parity tests needed to demonstrate disparity at a 95 percent confidence level.

Q. WHY DOES AMERITECH ILLINOIS USE A 95 PERCENT CONFIDENCE LEVEL?

A. The 95 percent confidence level is commonly used often in the field of statistical science. A 99 percent confidence level would have been a valid choice as well: It is also used frequently in scientific and statistical research. AT&T first suggested using the 95 percent confidence level for performance testing in 1998, during the FCC rulemaking on performance measurement.

As with the modified Z-test, other carriers quickly formed a consensus that 95 percent confidence would be appropriate. Although Ameritech Illinois believes that the 99 percent confidence interval would be appropriate, Ameritech Illinois accepted the 95

percent confidence interval in the spirit of compromise even though it would lead to an increased number of false findings of disparity. The FCC then approved the 95 percent confidence level when it approved the application of Bell Atlantic (now Verizon) to provide long-distance service in New York.

The 95 percent confidence interval has also been adopted by KPMG Consulting, which is conducting statistical tests of performance as part of its independent audit of OSS in Illinois and throughout the region.

Q. HOW DOES THE TABLE OF CRITICAL Z-VALUES IN THE AMERITECH ILLINOIS REMEDY PLAN WORK?

A. The table that appears in Section 9.3 of the Ameritech Plan is reproduced below. Table 1 lists the Z-values and number of apparently disparate tests results ("k") that would be needed to demonstrate disparity at the 95 percent confidence interval. It is based on a table developed by AT&T.[2]

[2] Affidavit of Colin Mallows, CC Docket No. 98-56 (Attachment C), pp. 18-19. The k-table within the Ameritech Illinois plan differs slightly from the one developed by AT&T. It is my understanding that Ameritech Illinois would be willing to alter the k-table in the Remedy Plan to make it consistent with the one developed by AT&T.

Table 1. Critical Z - Statistic Table

Number of Performance Tests	K Values	Critical Z-value
1	0	1.65
2	0	1.96
3	0	2.12
4	0	2.23
5	0	2.32
6	0	2.39
7	0	2.44
8	1	1.69
9	1	1.74
10-19	1	1.79
20-29	2	1.73
30-39	3	1.68
40-49	3	1.81
50-59	4	1.75
60-69	5	1.7
70 –79	6	1.68
80 – 89	6	1.74
90 – 99	7	1.71
100 – 109	8	1.68
110 –119	9	1.7
120 – 139	10	1.72
140 – 159	12	1.68
160 – 179	13	1.69
180 – 199	14	1.7
200 – 249	17	1.7
250 – 299	20	1.7
300 – 399	26	1.7
400 – 499	32	1.7
500 – 599	38	1.72
600 – 699	44	1.72
700 – 799	49	1.73
800 – 899	55	1.75
900 – 999	60	1.77
1000 and above	Calculated for Type-1 Error Probability of 5 percent	Calculated for Type-1 Error Probability of 5 percent

The first column of Table 1, lists the possible numbers of performance tests, while the third column lists the critical Z-value that applies to that number of tests. All you need to do is take the number of performance tests for the CLEC in question with at least 10 observations, find the applicable row in the table, then go over to the "Critical Z" column and find the applicable value. For example, if 19 performance tests apply to CLEC "1", the critical Z-value would be 1.79. If the number of performance tests for CLEC "2" is between 600 and 699, the critical Z-value would be 1.72. The math is based on standard statistical formulas (used by AT&T), and has already been done and recorded on the table.

After you take the critical Z- value from the table, you compare the actual Z-statistic for each performance test in question. If the Z-statistic is lower than the critical Z- value, the difference between wholesale and retail performance is not large enough to suggest disparity with 95 percent confidence, and we move on to the next test. The difference is not "statistically significant." If the number of recorded tests with Z- values larger than the critical Z exceeds the value of k in the third column of the same row, we can conclude with 95 percent confidence that there was disparity of service.

Q. CAN YOU PROVIDE SOME EXAMPLES OF HOW THE Z-TEST WORKS?

A. Yes. In December 2000 (from simulated data), the performance results for CLEC "174" showed activity in 76 performance measurement categories. Accordingly, based on the table of critical values at pages 11-12 of the remedy plan, the critical Z value for that CLEC was 1.68. The data for CLEC "174" included the following results for performance measure 27:

1) Mean Installation Interval - POTS - Residential - Fieldwork (Days), Chicago region

CLEC mean: 3.00 days ILEC mean: 4.87 days Z value: -6.09 Critical Z: 1.68

2) Mean Installation Interval - POTS - Business - No Fieldwork (Days), Chicago region

CLEC mean: 0.88 days ILEC mean: 0.30 days Z value: 1.56 Critical Z: 1.68

3) Mean Installation Interval - POTS – Residential - Feildwork (Days), IL North Central

CLEC mean: 4.00 days ILEC mean: 3.55 days Z value: 3.37 Critical Z: 1.68

For the first measure, the CLEC's results were better than retail. Because the remedy plan is only concerned with a disparity that goes *against* the CLEC, no further analysis is performed.

For the second measure, CLEC installations took more time than retail. The Z statistic, however, was only 1.56, less than the critical z value of 1.68. As a result, we conclude that the difference is due to random variation, and no further analysis is performed.

For the third measure, the Z-statistic of 3.37 exceeds the critical Z value of 1.68. This result suggests disparity, but with 95 percent confidence (or a 5 percent risk of error). Once all of the other parity tests for that CLEC have been performed in the month we can determine if the overall performance for that CLEC suggests that there is parity of service. Given that the CLEC has 76 performance metrics required for testing the k value drawn from Table 1 for this CLEC will be six. If more than six of these tests have a Z-

value exceeding 1.68 then we would be able to state that there is evidence of disparity based on a 95 percent confidence interval.

Q. WHAT IS THE PURPOSE OF THE THIRD COLUMN OF THE TABLE 1?

A. When thousands of statistical tests are performed, and each has a 5 percent Type I error rate, as is the case with performance testing, it is virtually guaranteed that large numbers of tests will appear to show disparity even when service is in perfect parity. The k-value in Table 1, in combination with the critical Z- value in that same table, establishes the number of apparent failures that would be needed to show a real disparity with the 95 percent confidence level the CLECs demanded.

Q. IS THE K VALUE EXACTLY 5 PERCENT OF THE NUMBER OF TESTS?

A. No. To say the individual statistical tests yield Type I errors 5 percent of the time *on average*, is the same thing as saying that the rate of flipping a fair coin and getting "heads" is 50 percent on average. For smaller sample sizes (say 30 flips), the actual number of heads will frequently vary quite a bit from that 50 percent average. We would expect that about half the time it will be higher and half the time lower. Similarly, given a Type I error rate of 5 percent, the number of false alarms for a given CLEC will likely exceed 5 percent half the time. Thus, setting "k" at exactly five percent is not statistically valid: It would give you only 50 percent "confidence" in the result.

To achieve the standard level of confidence, 95 percent, k is set slightly higher than 5 percent. As one would expect, where the number of measurement categories is small, the number of measures excluded is slightly higher than 5 percent: hence, eight categories would be excluded if 100 categories had data. But as the number of categories

increases – and there are now several thousand measurement categories – the value of k does approach 5 percent.

Q. IF THE K VALUE SAYS THAT SOME, BUT NOT ALL, OF THE APPARENT DISPARITIES ARE DUE TO RANDOM CHANCE, WHICH OF THOSE DISPARITIES ARE USED FOR ASSESSING REMEDIES?

A. As Mr. Fioretti shows in his affidavit, the remedy amounts for each performance tests reflect the importance of the related performance measure tested: either "high," "medium" or "low" priority. The K table is applied to the low priority measures first. Thus, Ameritech Illinois will pay on the highest-priority measures, which generally have the potential for highest remedies.

2. Parity Tests: Sample Sizes Less Than 30

Q. DOES AMERITECH ILLINOIS USE THE Z-TEST FOR ALL "PARITY" TESTS?

A. No. As I mentioned earlier, the Z-test is unlikely to function well on small sample sizes, such as less than 30 observations. Therefore, Ameritech Illinois uses alternative tests for sample sizes less than 30.

Q. WHY DOESN'T THE Z-TEST WORK ON SAMPLE SIZES UNDER 30?

A. Typically, as you look at more and more individual pieces of data or observations, they tend to fall into a pattern or distribution. Experience has shown that frequently once you look at 30 observations, the distribution of averages drawn from the population should be "normal," a statistical term that means the distribution is bell-shaped with about two-thirds of all observations falling within one standard deviation of the average. (This

tendency of the sample averages toward normality results from what statisticians call the "Central Limit Theorem.") The Z-test is designed to work well on these normal distributions. But if there are fewer than 30 observations, the sample averages are often fairly different from normal. And in this situation other statistical tests tend to be more effective and appropriate to use. For example, if you randomly pick only five repairs out of the universe of all repairs, you might end up picking the rare cases that are far away from the average (say, the repair that took twice as long as normal due to unusual weather or traffic), and thus do not fairly reflect the population of repairs as a whole. In those cases, the critical Z-value may not reflect the true cut-off for the 95 percent confidence interval.

Q. HOW DOES AMERITECH ILLINOIS ADDRESS THESE SITUATIONS?

A. The remedy plan uses alternative statistical tests known as "permutation tests," which are not sensitive to the non-normal distributions of the sample means that are often seen in sample sizes of less than 30 observations. In fact, permutation tests are always at least as good as Z-tests no matter how large the sample size; however, in large sample sizes they require a large number of repetitive calculations. Because the greatest advantage of the permutation test compared to the Z-test is observed when distributions are non-normal, Ameritech Illinois agreed with the consensus that formed around the practice of using permutation tests only when sample sizes were less than 30.

3. "Benchmark" Tests

Q. THE Z-TEST AND PERMUTATION TESTS ARE APPLIED TO "PARITY" TESTS. DOES AMERITECH ILLINOIS ADDRESS RANDOM VARIATION IN "BENCHMARK" TESTS?

A. Yes. In a benchmark test, wholesale performance is compared to a specified target. While the target itself doesn't move, wholesale performance is still subject to random variation, whatever you choose to compare against it.

Q. HOW DOES RANDOM VARIATION AFFECT BENCHMARK TESTS?

A. Let's use the return of mechanized completions as an example. The benchmark for this measure is 99 percent returned within 1 hour. Not every completion takes exactly 1 hour. And even if Ameritech Illinois returns 99 percent within 1 hour for all CLECs on average, you would expect the results for individual CLECs to vary: for half the CLECs, performance would be higher than 99 percent, but for the other half, performance would be less than 99 percent. Ameritech Illinois does not get credit for the half that are a little higher than the benchmark, and it should not be penalized for the half that are a little lower.

Q. HOW DOES AMERITECH ILLINOIS ADDRESS RANDOM VARIATION IN BENCHMARK TESTS?

A. The method is the same as that used by Southwestern Bell, and approved by the Texas commission and the FCC, in Texas. It does not use statistical analysis. It is my understanding that there is a buffer calculated for the benchmark as a compromise in Texas to reflect the random variations that occurs with all performance metrics.

B. <u>Analysis Of Disparity</u>

Q. IF THE STATISTICAL TESTS YOU DESCRIBE SHOW A DISPARITY, DOES THAT PROVE DISCRIMINATION BY AMERITECH ILLINOIS?

A. Not at all. First, as I described above, the statistical tests are designed to achieve only 95 percent confidence that disparity exists even when the statistical tests indicate there is disparity.

More importantly, these statistical analyses only indicate a *numerical* discrepancy between wholesale performance and the applicable standard (either retail performance or the benchmark). They do not identify the cause of that discrepancy. The discrepancy might have been the fault of Ameritech Illinois, but it might also be the result of something outside of Ameritech Illinois' control, such as an Act of God, or some mistake or misconduct by the CLEC or a third party (for example, a technician hired by the CLEC to coordinate a loop cut-over with Ameritech Illinois).

Q. CAN YOU GIVE US SOME EXAMPLES OF DISCREPANCIES CAUSED BY FACTORS OUTSIDE OF AMERITECH ILLINOIS' CONTROL?

A. Certainly. Let's assume that Ameritech Illinois and a CLEC each experience 100 "trouble reports" in January, and that the mean time to repair for the Ameritech Illinois customers was 3.4 hours, while the CLEC's customers experienced an average time of 5.0 hours. Statistical analysis might conclude there was a disparity, with 95 percent confidence. But further analysis might show that the time to restore service was exactly 7 hours during the first week (due to a blizzard) for all carriers, and 3 hours the rest of the month (again, for all carriers). It might also show that 50 percent of the CLEC customers reported trouble during the blizzard week, while only 10 percent of the Ameritech Illinois customers reported trouble that week. Thus, the CLEC average would be 5 hours (50 percent at 3 hours, and 50 percent at 7 hours) while the Ameritech Illinois average would be 3.4 hours (90 percent at 3 hours, and 10 percent at 7 hours). The discrepancy would

not be due to any wrongdoing by Ameritech Illinois, but would result from the fact that a higher percentage of CLEC repairs happened to coincide with the blizzard.

Q. HOW DOES THE AMERITECH ILLINOIS REMEDY PLAN ADDRESS THESE POSSIBILITIES?

A. The remedy plan takes a pro-CLEC approach. It presumes that the discrepancy was the fault of Ameritech Illinois and requires Ameritech Illinois to pay a remedy unless the Commission finds that a remedy is not warranted. The plan then establishes an expedited procedure for Ameritech Illinois to seek a waiver from the Commission. If Ameritech Illinois does not initiate that procedure before the date remedy payments are due, the pro-CLEC presumption stands and Ameritech Illinois must pay the remedy. The waiver procedure is described in more detail in the affidavit of Mr. Fioretti.

Q. IF THE STATISTICAL TESTS YOU DESCRIBE SHOW DISPARITY DOES THAT MEAN THAT DIFFERENCES IN SERVICE ARE LARGE OR NOTICABLE TO CUSTOMERS?

A. No. These statistical tests are designed to determine if there is enough empirical evidence to establish some disparity no matter how large or small the disparity. A one-minute disparity in installing POTS between wholesale and retail, for example, would probably not affect consumer decisions or even noticed by consumers. The tests used by Ameritech Illinois do test for these small differences and assess remedies based on them, but they by no means imply that these differences are large enough to be relevant to consumers or competition. In this way, Ameritech Illinois' statistical tests and remedy plan are pro-CLEC, paying remedies in cases where there is a statistical difference even if the difference is too small to be relevant to consumers.

C. Type I and Type II Errors

Q. EARLIER, YOU STATED THAT THE STATISTICAL TESTS ARE DESIGNED FOR A TYPE I ERROR RATE OF 5 PERCENT. CAN YOU ILLUSTRATE WHAT THIS MEANS?

A. Yes. A Type I error is the risk that random variation will lead you to decide there is a disparity when, in fact, there is parity. To return to the analogy of the coin flip, a fair coin is one that has equal probability of turning up heads and tails. If you flip it 50 times, the expectation is that you will get heads 25 times and tails 25 times. But in fact this rarely happens. The laws of probability show that there is only about an 11 percent chance that a fair coin flipped 50 times will produce exactly 25 heads. So if you concluded there was "parity" only when you saw 50 heads, you would make the right call only 11 percent of the time. The other 89 percent of the time, you would see something other than 25 heads, and you would mistakenly think there was some "disparity" with the coin one way or the other. Declaring this fair coin to be biased would be a Type I error.

The Ameritech Illinois plan uses a different and more accurate decision rule. As I described above, the plan establishes a "critical value," at the point where there is a 95 percent certainty that disparity exists. In this example, we know from the laws of probability that there is about a 5 percent chance that 50 tosses of a fair coin will result in 32 or more heads. If our 50-flip test gives us the critical value of 32 heads or more, the plan presumes that there is a disparity. (The plan is not concerned with disparities that favor the CLEC, so here we are not concerned with a coin that yields too many tails.) That conclusion would be wrong about 5 percent of the time (because even a fair coin

generates 32 or more heads about 5 percent of the time), so the Type I error rate is 5 percent.

Q. YOU ALSO REFERRED TO A TERM CALLED "TYPE II" ERRORS. CAN YOU ELABORATE?

A. Yes. A Type II error occurs when you mistakenly decide there is parity, when in reality there is disparity. A coin might be unfairly biased such that it comes up "heads" more than half the time on average – say, 70 percent of the time. But because of random variation, you might still see 25 heads and 25 tails when you flip it 50 times, leading you to decide the coin was fair. That would be a Type II error: a false "pass" (in contrast to a Type I error, which is a false failure).

Q. IS THERE A WAY TO MEASURE THE RATE OF TYPE II ERRORS IN PERFORMANCE TESTING?

A. Not with the information that is currently available. As with grading any other test, to decide whether a statistical test yields the right answer or an error, one needs to define what the right answer is. If a test suggests disparity, you need to know what parity is to determine the probability of a Type I error. That is simple, because "parity" or equality is a straightforward concept that is easy to define and agree upon. By contrast, if a test suggests parity, the only way to know the probability of a Type II error is to define *dis*parity. A small difference is harder to detect and thus easier to miss. To follow our analogy, there might be some minute imperfection in a coin that leads it to turn up heads 50.000000001 percent of the time, but that difference is not material, and failing to detect it would not be cause for concern.

Given that performance testing is being performed to promote competition in the telecommunications market, it makes sense to say that a material difference in performance is one that would be large enough to alter consumer behavior in a way that would affect competition. If a difference in performance is too small to alter consumer behavior or competition, failure to identify a disparity of this size (a Type II error) is not relevant for the purposes of the remedy plan.

Q. HOW WOULD ONE GO ABOUT DEFINING MATERIALITY?

A. Obviously, materiality will vary depending on the measure and product being considered. It may also depend on other factors that vary by geography or potentially by a range of additional variables. The CLECs themselves agree, and suggest that telephony experts, not statisticians, should be consulted to decide what constitutes a substantial or material difference:[3]

> While statistical science can be used to evaluate the impact of different choices of these parameters [δ_j, and the alternative hypothesis among others], there is not much that an appeal to statistical principles can offer in directing specific choices. Specific choices [of δ_j] are best left to telephony experts.[4]

The CLEC proposal goes on to say that "[t]he bottom line here is that beyond a few general considerations, like those given above, a principled approach to the choice of the alternative hypotheses to guard against must come from elsewhere."[5] Because the definition of a material difference could depend on the measure, geographies, and on sample sizes,[6] you would have to do extensive research to find out what difference in service provided to retail and wholesale customers would have a material impact on

---

3 *CLEC Proposed Remedy Plan for Illinois*, filed March 12, 2001, p. 27.
4 *Id.*
5 *Id.* p. 28.
6 *Id.*

competition. That would in turn require a vast array of information about how customers react to differences in levels of service. At this time, such information does not exist, and it certainly does not exist for each performance measure, local telecommunications product, or geographical disaggregation, let alone for every relevant combination that is included in the remedy plan.

Q. GIVEN THAT THERE IS NO PRACTICAL WAY TO DEFINE "DISPARITY," HOW CAN ONE ADDRESS THE RISK OF TYPE II ERROR?

A. Although there is no way to precisely measure and control Type II error in this context, we do know enough about its tendencies to reasonably address it. First, we know that Type II errors become less likely as the disparity becomes larger (because it is easier to detect a large difference than a small one), or as the number of transactions increases (because there is more information to include in the statistical analysis). That is helpful, because we are most concerned with disparities that are large, or that affect a large number of customers.

Q. HOW DOES AMERITECH ILLINOIS' REMEDY PLAN APPLY THESE PRINCIPLES?

A. Ameritech Illinois' plan is designed to use statistical methods that are generally accepted in the scientific community. It starts by conducting statistical analysis on all performance metrics.. If differences in service appear that meet or exceed a critical Z-value and for a number of disparities listed in the K table, they are automatically presumed to reflect true underlying disparity, even though there may be a valid reason for the observed difference as well as a 5 percent chance that the difference is due to random variation (set at 5 percent) that there is really parity. This provides sufficient assurance that a substantial

disparity will be detected. As the observed wholesale performance gets farther from retail, it is more likely that there is some true underlying disparity; conversely, as observed wholesale performance gets closer to (or even better than) retail, there is less likelihood of an underlying disparity.

Q. WHO DEVELOPED THIS APPROACH TO TYPE I AND TYPE II ERROR FOUND IN THE AMERITECH ILLINOIS PROPOSAL?

A. It was first developed by AT&T, in the FCC's docket on performance measurement. AT&T's expert stated in his affidavit that "AT&T proposes to set the Type I error at no more than the conventional level of 5 percent." As he explained, "[t]his controls the frequency of false alarms to be at most 5 percent while making the probability of Type II errors small for violations that are of substantial size. Using a one-tailed test for Type I error at about the 5 percent level thus strikes a reasonable balance."

As with the modified Z-test, a consensus formed around this methodology. Thus, Bell Atlantic adopted the same approach for its remedy plan in New York. The FCC endorsed this approach when it approved Bell Atlantic's application to provide long-distance service in New York.

III. COMPARISON TO CLECS' PROPOSED METHODOLOGY

Q. CAN YOU SUMMARIZE THE DIFFERENCES BETWEEN THE STATISTICAL METHODOLOGY OF THE AMERITECH ILLINOIS REMEDY PLAN, AND THE PLAN PROPOSED BY THE CLECS?

A. Yes. As I noted above, Ameritech Illinois' remedy plan reflects a consensus that was achieved between incumbent LECs, CLECs, state commissions, and the FCC regarding

generally accepted statistical principles and techniques. The present CLEC plan, however, abandons that consensus. The main differences between the two plans are:

(A) **Type I and Type II Errors:** While the Ameritech Illinois plan uses the conventional 95 percent confidence level, the CLEC proposal attempts to balance between the two types of errors.

(B) **Small Sample Sizes:** Where there are less than 30 transactions (which we have found happens often), Ameritech Illinois uses tests designed to work on small samples. The CLEC plan uses tests that are *not* designed for small samples.

(C) **Missing Pieces:** The CLEC plan contains a large number of missing parameter values, which prevents the reader from fully understanding it or implementing it in practice. The CLEC plan has many other missing features such as tables and test definitions. The fact that the CLEC plan is not fully developed is evident from the fact that when AT&T was asked to provide programs and present simulated data following their plan they did not, or were not able to, follow the written plan that they presented to the ICC.

A. <u>"Balancing" Type I and Type II Errors</u>

Q. HOW DOES THE CLEC PROPOSAL DIFFER FROM AMERITECH ILLINOIS' PLAN WITH RESPECT TO TYPE I AND TYPE II ERRORS?

A. As I described above, Ameritech Illinois controls Type I errors by setting the risk of error at 5 percent. This has the effect of addressing Type II errors as well, and is designed to

achieve a fair compromise: reducing the risk that remedies will be assessed in error (Type I error) while reducing the risk that a large disparity would go undetected (Type II error). The CLEC plan, however, purports to make the risk of Type I and Type II errors exactly equal. As I discuss later, the "Type I – Type II balancing" proposed by the CLECs does not provide balanced remedies. When it comes to the incentive structures that they create, the CLEC plan is very unbalanced, creating hundreds of millions of dollars of remedy payments even when Ameritech Illinois is providing service that is in parity.

Q. IS IT FEASIBLE TO PRECISELY "BALANCE" TYPE I AND TYPE II ERRORS IN THAT MANNER?

A. Not the way the CLECs have developed their plan. As I testified earlier, you cannot measure Type II error without first defining the level of disparity you are looking for. That would require extensive study of the impact differing levels of performance have on each product, service, and geographical area. At present the CLECs have not provided one and as far as I know no such analysis exists.

Q. HOW DOES THE DEFINITION OF DISPARITY AFFECT THE MEASUREMENT OF TYPE II ERROR?

A. The definition of disparity has a dramatic effect on the estimated Type II error, so an accurate definition would be critical to make the CLEC balancing plan work. For a given sample size, it is easier to miss a small disparity than a large one, so the risk of Type II error is higher. Even the CLEC proposal states that the definition of disparity will greatly

influence the outcome of the tests, and that "even small disagreements among experts" in the choice of that definition "could be very important."[7]

Q. HOW DOES THE CLEC PROPOSAL ADDRESS THIS ISSUE?

A. It doesn't. Instead, the CLEC proposal simply selects an arbitrary fixed definition of materiality that it applies to all performance measures and categories. For continuous or duration measures, it defines a material disparity at 0.25 of the population standard deviation of the retail observations for the measure being tested. This value of 0.25 is called δ_j ("delta") in their proposal and it is the same for every single test. The CLEC proposal does not justify this choice other than to say that the CLECs agreed to a "joint proposal as an opportunity to study the impact of the 0.25 delta pending the six month review of the plan."[8]

Q. IS THAT A SATISFACTORY SOLUTION?

A. Absolutely not. The CLECs' arbitrary definition of disparity is clearly an unsatisfactory method for defining an element that is of such vital interest in determining the outcome of performance tests. And the CLECs themselves agree. The CLEC proposal says that a fixed value of δ_j "does not seem sensible."[9]

Q. WHAT EFFECT WOULD THE CLEC BALANCING PROPOSAL HAVE?

A. In most samples encountered in the remedy plan it will raise the Type I error rate above 5 percent. To illustrate, assume we did a performance test on a sample of 50 coin flips. Using the conventional 5 percent Type I error rate, the Ameritech Illinois plan would set

---

7 *CLEC Proposed Remedy Plan for Illinois,* p. 28.
8 *Id.* p. 13.
9 *Id.* p. 28.

the critical value at 32, and would conclude that the coin was unfair if heads came up 32 or more times. But the CLEC proposal would say that 29 heads or greater would indicate disparity, raising the Type I error rate to 16 percent.

Q. WHAT IS THE EMPIRICAL SUPPORT FOR SETTING THE MEASURE OF MATERIAL DIFFERENCE AT 0.25?

A. There isn't any. I cannot stress enough that the CLECs' underlying assumption (that a difference of .25 standard deviations is material enough that failing to assess remedies would be a Type II error) is completely arbitrary. There is no empirical evidence about consumer behavior to support it, and I have not seen anyone present any evidence about consumer reaction to differences in service to support it in any of the proceedings that I have attended or reviewed. The CLECs are trying to "balance" Type I error, which is well defined, against a random number that they call Type II error. One could just as easily say that a difference is not material unless it is 1 or 2, or more, standard deviations.

Q. WHAT WOULD BE THE PRACTICAL EFFECT OF INCREASING THE TYPE I ERROR RATE IN THE WAY THE CLEC PROPOSAL WOULD DO?

A. For instance, assume that all the performance tests had a Type I error rate of 30 percent. Under this scenario, even if Ameritech's underlying performance was in perfect compliance there would still be a finding of disparity - and Ameritech Illinois would still pay remedies – on 30 percent of the performance tests. If there were 5,000 total performance tests, Ameritech Illinois would pay 1,500 remedies based purely on testing error. And if the average remedy were $20,000, Ameritech Illinois would pay $30 million in remedies where it should have paid none. As I show below, the amount of

undeserved remedies under the CLEC plan would work out to $28 million per month even when Ameritech Illinois provided service that was in perfect parity.

Thus, there are two possible outcomes. One would be that Ameritech Illinois would continue to provide perfectly fair or nondiscriminatory service, but still pay CLECs $28 million per month in undeserved remedies. The CLECs would have their cake and eat it too: benefiting in the market from nondiscriminatory performance, and benefiting at the bank from undeserved remedies. The other possible outcome would be that Ameritech Illinois would try to avoid remedies by discriminating in favor of CLECs. That goes against the principles of parity and balance that the remedy plan is supposed to enforce.

Q. BESIDES δ ARE THERE OTHER DEFINITIONS OF MATERIALITY THAT THE CLECS INCLUDE IN THEIR PROPOSAL?

A. Yes, the CLECs have introduced two additional measures of materiality, ε_j and φ_j, which are used to determine the material difference for rate and proportion measures, respectively.[10] The CLECs provide no indication of what the values of ε and φ should be. Without some definition of the values of ε_j and φ_j, the Type I –Type II balancing tests that the CLECs propose cannot be calculated because there is no alternative hypothesis (*i.e.*, no definition of disparity) to test against. Given that rates and proportions compose approximately two-thirds of all performance measures tested, this omission makes the CLEC proposal literally inoperable.

10 *CLEC Proposed Remedy Plan for Illinois*, Attachment 2, pp. 4-6.